BIG DATA MEETS LINKED DATA – WHAT ARE THE OPPORTUNITIES?

Jon Blower [1], Maik Riechert [1], Nino Pace [2], Manolis Koubarakis [3]

1. Department of Meteorology, University of Reading, United Kingdom
2. Advanced Computer Systems, Rome, Italy
3. National and Kapodistrian University of Athens, Greece

ABSTRACT

The worlds of "Big Data" and "Linked Data" are have evolved somewhat separately but both are highly relevant to the development of systems for sharing and processing large volumes of Earth Observation data. This paper focuses on Linked Data, examining this field from a user-centric point of view, based on research carried out in a number of recent European projects. A central theme is that Linked Data can help to open up Earth Observation data to new users and communities.

Index Terms— Earth Observation, Data discovery, Data management, Semantics, Quality, Visualization

1. BIG DATA AND LINKED DATA

The term "Big Data" has numerous definitions but practitioners recognize many challenging aspects, including data *volume* (e.g. where data become too large to process using traditional techniques), data *velocity* (e.g. where data are captured at high rates from sensors), data *variety* (e.g. bringing together data from multiple sources) and data *veracity* (i.e. data quality).

All of these challenges are faced by users of Earth Observation (EO) data. Whereas questions of volume and velocity can be addressed through the development of high-performance "Big Data" capture and processing systems (e.g. http://www.jasmin.ac.uk), questions of variety and veracity can only be addressed by looking more carefully at how we can help users by *describing* data better and by sharing this information more effectively within and outside the EO community. Linked Data techniques provide a number of ways to achieve this. Therefore, Linked Data and Big Data are entirely complementary ideas.

There are many possible views of the concept of "Linked Data". At a high level, the promise of Linked Data is that we will be able to use the full power of the Web to break down barriers between communities that were hitherto separated. The key features of Linked Data are:

1. Information is *decentralized* (like the Web in general). Anybody can say Anything about Any topic (this is the "AAA" principle).
2. Information is *machine-readable* (unlike the wider Web).
3. Through wide use of certain *standards* (see below), diverse sources of information can be linked together unambiguously.

In this way, we do not have to collect all the possible information about a topic into a single place; we can use the Web as a highly-distributed, decentralized research environment. This allows information providers to focus on publishing the information that they "own" and in which they are expert.

The key technical steps that an information provider must follow in order to participate in the Linked Data web are [1]:

1. Create (or, preferably, reuse existing) unique and persistent identifiers for the important "things" in a community (e.g. datasets, publications, algorithms, instruments).
2. Allow users to "look up" these identifiers on the web to find out more information about them (in other words, the identifiers are essentially Web addresses, i.e. URLs).
3. Make this information machine-readable in a community-neutral format. RDF (Resource Description Framework) is the preferred choice, using terms from widely-agreed vocabularies.
4. Within this information, embed links to other things and concepts and say *how* these are related.
5. Optionally, provide web service interfaces to allow the user to perform sophisticated queries over this information. SPARQL is the preferred query language for this.

By describing data more precisely and by using high-level, community-neutral standards like RDF and SPARQL (both are standards of the World Wide Web Consortium, W3C), information should become more widely accessible and understood by users in different fields. Perhaps the main challenge is to identify the most appropriate *vocabularies* (defined terms) to describe data. Interoperability is enhanced if the same vocabularies are widely adopted; however, frequently data providers find that they need to create new terms and identifiers to describe their data accurately. Designing and publishing these new terms can be a significant undertaking. An example of a highly-relevant RDF vocabulary is GeoSPARQL [2], which describes how to encode geospatial data in RDF and query them using SPARQL.

2. USE CASES FOR LINKED DATA IN EO

The technological view of Linked Data is well-established, as is the high-level view of the overall benefits that it can bring. But the concrete benefits to the EO user community are rarely described explicitly. In this section we briefly outline some promising use cases based on experience derived from numerous European research projects, notably TELEIOS (http://www.earthobservatory.eu/), CHARMe (http://www.charme.org.uk, [3]), MELODIES (http://melodiesproject.eu, [4]) and LEO (http://linkedeodata.eu [5]).

2.1. Enabling data discovery

In the past years, considerable effort has been devoted to the development of special-purpose catalogue and search infrastructure to enable discovery of scientific data, such as the Global Earth Observation System of Systems (GEOSS).

To complement these initiatives, Linked Data can also help users to find relevant information using standard Web mechanisms. One way is simply for the user (or automated system) to traverse the links between information sources as they would on the Web. Secondly, we believe that in the coming years it will become increasingly possible to discover EO data through "mass market" search engines such as Google and Bing. These search engines are adopting Linked Data techniques for harvesting structured (RDF) information about datasets from websites. Data providers can describe their datasets using an appropriate vocabulary (e.g. http://schema.org), and search engines can harvest this information. The potential is that search engines will be able to provide richer information about datasets they find, and automatically link this with other related information. The need for special-purpose, community-specific search engines may therefore reduce and EO data will become easier to discover by new users.

2.2. Helping users to understand data.

By enabling different sources of information to be linked together in a structured fashion, Linked Data can help users to understand much more about datasets. In particular, the interrelationships between a dataset and the information surrounding it (e.g. upstream data, processing chains, publications, experts) can be described quite naturally in a Linked Data form and discovered by users (see Figure 1). User-supplied annotations (see section 2.4) can also play a role here.

2.3. Dynamically combining and integrating data

Currently the process of combining multiple data sources from different communities (e.g. in a web application) is usually lengthy and dominated by low-level technical concerns of data reformatting. The use of RDF as a universal "lingua franca" and SPARQL as a standard web service interface can greatly help developers of such systems to bring data together *at runtime* from multiple sources, without the need for further data manipulation, harmonization or conversion. Now that the essential tools are maturing (see section 3), such applications are beginning to emerge (e.g. http://almere.pilod.nl/bgtld/v2).

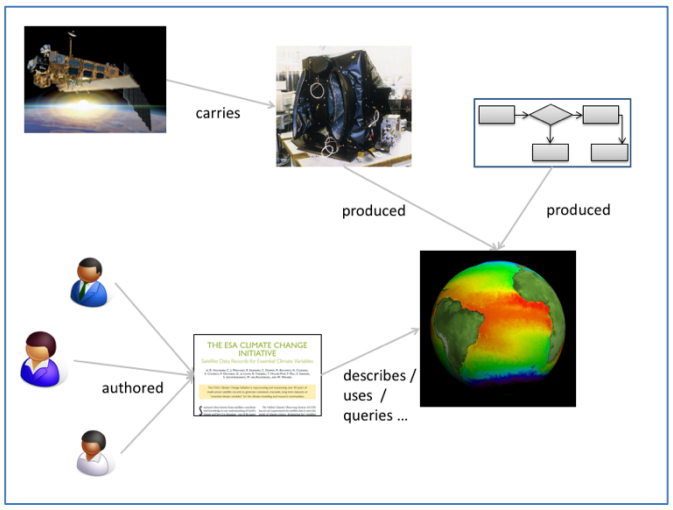

Figure 1: A simplified illustration of the links between EO platforms, instruments, algorithms, datasets, publications and scientists, represented as Linked Data.

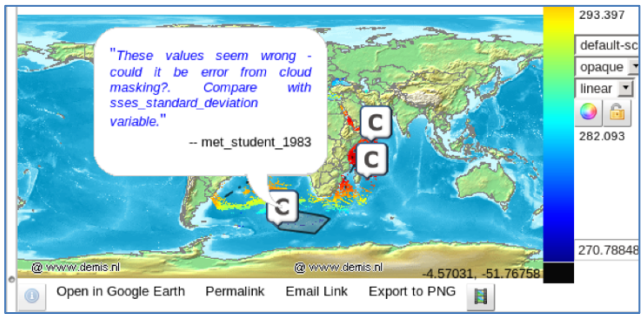

Figure 2: The "CHARMe Maps" tool, which allows user-supplied annotations to be linked with datasets and even specific subsets of datasets. Each annotation is represented by a "C" icon.

2.4. Allowing users to share information with each other

Currently users of EO data receive most of their information from the original provider of the data. However, the provider usually does not usually have complete information about important topics such as data quality and fitness for purpose; such information is often gathered or supplemented by the user community. The CHARMe project developed a system, based upon Linked Data, to allow users to create and share *annotations* about EO data [3]. These annotations include free-text comments, links to relevant publications, statements about data quality and more (Figure 2). The

system exploits the fact that data providers are increasingly assigning persistent identifiers (e.g. DOIs) to their data. These identifiers act as "anchors" for the annotations, so that users can be very clear about what they are annotating, be it a dataset, a single satellite image, a sensor or even another annotation.

3. NEW TOOLS AND APPROACHES

Data structures and services such as RDF and SPARQL are not familiar to most users of Earth Observation data. This section describes a suite of recently-developed open-source tools that help data providers and users.

Strabon [6] is an RDF data store, that is capable of storing geospatial Linked Data that changes over time. Data can be queried using both the OGC-standard GeoSPARQL dialect [2] and stSPARQL. Strabon supports a wide range of geographic functionality, such as support for spatial geometries and a wide range of coordinate reference systems. Strabon can be used to model temporal domains and concepts such as events and facts that change over time.

GeoTriples [7] is an open source tool for converting geospatial data from several common formats into RDF. It is able to generate and process mappings from source files (GML, KML, GeoJSON, ESRI Shapefile, CSV) and spatially-enabled relational databases to RDF graphs. GeoTriples employs by default the well-known ontologies like GeoSPARQL and stSPARQL, without being tightly coupled to a specific vocabulary. It offers rich support for processing geospatial data and is able to scale to large data volumes.

Ontop-spatial (https://github.com/constantB) is a geospatial extension of the system Ontop (http://ontop.inf.unibz.it). In contrast with Strabon, Ontop-spatial does not require that data be converted stored in a special RDF data store. Instead data remain in their source format (e.g. relational databases or shapefiles), and Ontop-spatial creates a "virtual" RDF graph that can be queried efficiently using SPARQL and GeoSPARQL. Therefore this system is suitable for data providers who prefer not to convert their data to RDF.

The **Silk** Link Discovery Framework (http://silkframework.org/) can be used in order to discover relationships between entities in the datasets. In particular, using Silk one can pre-process the data and create spatial, temporal and other types of relationships between data items. This can greatly increase the efficiency of complex queries.

Sextant (http://sextant.di.uoa.gr/) is a web- and mobile-based application for exploring, interacting and visualizing time-evolving linked geospatial data [8]. The functionalities of Sextant include the exploration and visualization of linked spatiotemporal data, the creation, sharing, searching and collaborative editing of maps and the production of statistical charts. While the tool heavily utilizes semantic web technologies, Sextant provides a user-friendly interface (Figure 3) to allow both domain experts and non-experts to use all features provided. Sextant can ingest heterogeneous data from files and web services that support a variety of standards, including KML, SPARQL and GeoSPARQL.

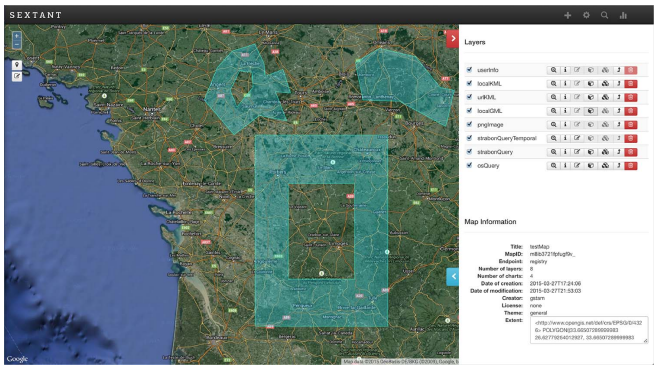

Figure 3: Screenshot of Sextant, showing visual integration of Linked Data from a number of sources.

The technologies we have described so far mainly relate to vector data, i.e. data about discrete geospatial features such as the locations of in situ observations and the boundaries of satellite images. However, satellite image data themselves usually take the form of multidimensional arrays, which can be very large in size. It would not be efficient to convert all the individual pixels in large images to RDF structures. However, the metadata about images (e.g. concerning the variables that are measured and the instrument itself) is highly amenable to being described in RDF. Therefore the EO community needs a way to handle RDF and array data seamlessly.

CoverageJSON (http://tinyurl.com/covjson) is a new (currently experimental) lightweight format for encoding scientific data (including Earth Observation data) in the JSON format. JSON (JavaScript Object Notation) is the predominant format preferred by modern web developers for consuming data feeds in web-connected applications. A "coverage" is a data structure that maps positions in space and time to data values, and can be thought of as an overarching concept that encompasses many kinds of scientific data (see the ISO19123 standard).

CoverageJSON has two main goals: (1) to integrate multidimensional array data with detailed semantic information in the same format; and (2) to make it easier for web developers to ingest scientific data into interactive web applications for in-browser visualization and processing. The key to the format is the use of JSON-LD (http://json-ld.org, an RDF variant) for encoding semantic information, whilst using "plain" JSON for data elements such as arrays. This allows the format to be easy-to-use and efficient, whilst retaining the expressive power of RDF for those applications that require it. It is therefore intended to provide a bridge between the EO and Linked Data communities.

The format has been developed and tested against use cases within the MELODIES project: e.g. in-browser reclassification of land cover datasets (Figure 4), subsetting via user-supplied polygon shapes, derivation of statistics, and intercomparison of multiple datasets.

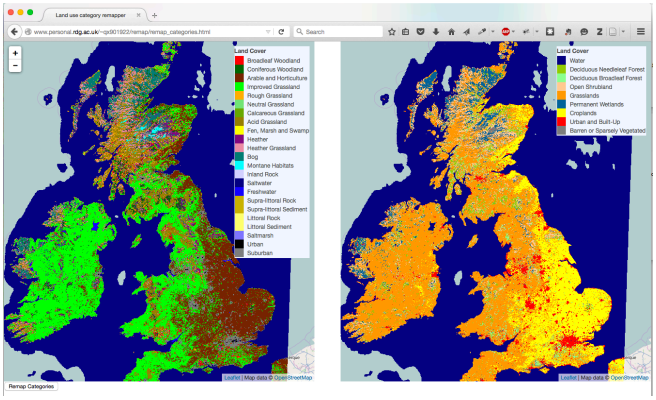

Figure 4: Reclassification of a land cover product from the MELODIES project from the MELODIES classification scheme (left) to the MODIS scheme (right). Data are loaded from a server in CoverageJSON format, then visualized and manipulated within the web browser.

4. CONCLUSIONS

In this paper we have discussed how Big Data and Linked Data are related concepts, contributing to advances in data discovery, understanding and analysis. A key current challenge is efficiency: RDF is extremely expressive but is not usually the most efficient format for data storage or analysis. Furthermore, we find that the conversion of data to RDF can be a burden on application developers, if the original data provider has not provided data in this way. Both of these issues are being addressed through the development of new tools and approaches, but it is clear that Linked Data techniques must be used *alongside* other more efficient "Big Data" techniques, depending on the user's need.

A key goal of Linked Data is to make data more widely accessible by a much wider community. We recommend that EO data providers and application developers consider the following activities to contribute towards this goal:

1. Publish "fundamental" information (such as authoritative information on datasets, instruments, satellites etc.) as Linked Data, so that the community can discover and link to it.
2. Consider publishing data and metadata in widely-used, "web-friendly" formats such as JSON(-LD), thereby making information more usable by typical developers who are not EO data specialists.
3. Publish structured metadata in forms that are understood by mass-market search engines, to enable easier data discovery.
4. Participate in joint efforts between the web community and the geospatial community in order to contribute to the standards that will bring these communities together. A current example of such an initiative is the joint OGC/W3C Spatial Data on the Web Working Group (http://tinyurl.com/sdwwg).

ACKNOWLEDGEMENTS

The TELEIOS, MELODIES, LEO and CHARMe projects have received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement numbers 257662, 603525, 611141 and 312641 respectively.

REFERENCES

[1] C. Bizer, T. Heath and T. Berners-Lee (2009) Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, Vol. 5(3), Pages 1-22.

[2] M. Perry and J. Harding (eds.) (2012) OGC GeoSPARQL - A Geographic Query Language for RDF Data v1.0 Open Geospatial Consortium document 11-052r4

[3] D. Clifford, R. Alegre, V. Bennett, J. Blower, C. DeLuca, P. Kershaw, C. Lynnes, C. Mattmann, R. Phipps, and I. Rozum (2015) Capturing and sharing our collective expertise on climate data: the CHARMe project. Bulletin of the American Meteorological Society. doi:10.1175/BAMS-D-14-00189.1

[4] J. Blower, D.Clifford, P. Gonçalves, and M. Koubarakis (2014) The MELODIES project: integrating diverse data using Linked Data and cloud computing. In: 2014 ESA conference on Big Data from Space (BiDS'14), Frascati, pp. 244-247.

[5] M. Koubarakis (2014) Linked Open Earth Observation Data: The LEO Project. Informal Proceedings of the Image Information Mining Conference: The Sentinels Era (ESA-EUSC-JRC 2014). Bucharest, Romania, 5-7 March.

[6] K. Kyzirakos, M. Karpathiotakis and M. Koubarakis (2012) Strabon: A Semantic Geospatial DBMS. In the 11th International Semantic Web Conference (ISWC 2012). Boston, USA.

[7] K. Kyzirakos, I. Vlachopoulos, D. Savva, S. Manegold, M. Koubarakis. GeoTriples (2014) A Tool for Publishing Geospatial Data as RDF Graphs Using R2RML Mappings. Terra Cognita 2014, 6th International Workshop on the Foundations, Technologies and Applications of the Geospatial Web, in conjuction with ISWC 2014. Riva del Garda, Trentino, Italy.

[8] K. Bereta, C. Nikolaou, M. Karpathiotakis, K. Kyzirakos, and M. Koubarakis. SexTant (2013) Visualizing Time-Evolving Linked Geospatial Data. In the 12th International Semantic Web Conference (ISWC 2013). Sydney, Australia. Demo paper.